文献与数据学报
第 5 卷 第 3 期 2023 年 9 月

基于知识图谱的黄河流域非遗资源智能问答研究

张 强[1, 2, 3] 吴艳飞[1] 高 颖[1, 2] 周树斌[1]
（1. 华中师范大学信息管理学院，武汉 430079；
2. 中国人民大学数字人文研究院，北京 100872；
3. 安徽师范大学新闻与传播学院，芜湖 241002）

摘 要：［目的 / 意义］利用数字技术赋能非遗资源的深度挖掘，揭示黄河流域非遗资源的关联关系，对黄河流域的非遗文化保护与传承具有重要意义。［方法 / 过程］以黄河流域非遗资源为研究对象，采用自顶向下的方式构建黄河流域非遗资源的知识图谱，并构建以用户交互为核心的智能问答系统。［结果 / 结论］本研究构建的基于知识图谱的黄河流域非遗资源智能问答系统实现了黄河流域非遗资源的多维知识发现，为非遗资源的相关研究提供了新的思路。

关键词：知识图谱 非物质文化遗产 黄河流域 智能问答
分类号：G254；TP391.3
DOI：10.31193/SSAP.J.ISSN.2096-6695.2023.03.09

黄河是中华民族的母亲河，孕育着中华民族的优秀传统文化，其流经范围广，空间跨度大，横跨中国九个省区，积淀了丰富多样的非物质文化遗产，包含民间文学、民间音乐、民间舞蹈、传统戏曲、传统体育、民间美术、传统技艺、传统医药、民俗等类别。2019 年，习近平总书记在《在黄河流域生态保护和高质量发展座谈会上的讲话》报告中指出，要推进黄河文化遗产的系统保护，守好老祖宗留给我们的宝贵遗产；要深入挖掘黄河文化蕴含的时代价值，讲好“黄河故事”，延续历史文脉，坚定文化自信，为实现中华民族伟大复兴的中国梦凝聚精神力量[1]。2021 年，中共中央办公厅国务院发布《关于进一步加强非物质文化遗产保护工作的意见》[2]，强调大力推动非物质文化遗产保护工作的重要性。黄河流域非物质文化遗产作为黄河文化的重要载体，

［作者简介］张强（ORCID：0000-0002-7020-8427），男，博士研究生，研究方向为数字人文，Email：zhangqiang_dh@163.com；吴艳飞（ORCID：0009-0009-1235-4467），女，硕士研究生，研究方向为数字人文，Email：323515425@qq.com（通讯作者）；高颖（ORCID：0000-0002-2496-6286），女，硕士研究生，研究方向为数字人文，Email：yinggao1213@163.com；周树斌（ORCID：0000-0002-9657-8178），男，博士研究生，研究方向为数字人文，Email：zshubin001@163.com。

推动黄河流域非物质文化遗产资源的研究、挖掘、保护、传承与融合，对弘扬黄河文化、传承黄河文化基因具有重要的研究价值。

数字人文作为信息技术与人文学科交叉衍生出的新兴领域，提供了分布协作性的跨学科研究范式[3]。随着数字技术的快速发展，人文学科与信息技术的关系愈发紧密、融合程度日益加深，为数字人文和新文科研究提供了发展空间[4-5]。知识图谱是数字人文研究中一项关键的知识组织技术，是一种结构化的语义知识网络，由节点和边构成，用于描述物质世界中的概念、实体及其相互关系[6]。借助知识图谱，可以揭示黄河流域非物质文化遗产资源中的实体及实体间的关系，挖掘资源间的隐含内容，以可视化的方式将实体及其关系展示出来。同时，运用知识图谱技术赋能黄河流域非遗资源，可以扩展黄河流域非遗资源的知识发现等服务，进而实现对黄河流域非遗资源的深度开发利用。

本研究以黄河流域非物质文化遗产资源作为研究对象，利用自顶向下的方式构建黄河流域非物质文化遗产资源知识图谱，实现了黄河流域非遗资源的关联组织与语义表示，最终构建出面向用户交互为核心的智能问答系统。

1 相关研究

1.1 文化遗产领域知识图谱构建研究

知识图谱是一种具有语义处理能力与开放互联能力的知识库，可以将结构松散、多源异质的数字资源组织起来，实现知识间的互联[7]。目前，知识图谱在文化遗产领域形成了一定的研究成果，代表性的工作主要有：赵雪芹等通过构建万里茶道数字资源本体，再基于该本体构建万里茶道数字资源知识图谱，实现了资源的知识关联与表示[8]；高劲松等以可移动文物为对象，构建可移动文物知识本体，然后基于映射规则将本体的实例映射到图数据库中来构建可移动文物知识图谱，最后以关联数据的形式发布出来[9]。在非物质文化遗产领域，代表性的工作有：范青等构建非物质文化遗产知识图谱模型，基于模型展示资源之间的关联关系，实现数据关联，为非遗知识数字化提供了新方案[10]；赵雪芹等利用领域知识图谱实现非遗档案资源的知识组织，并以“华县皮影”非遗档案资源为例来验证知识图谱在非遗档案中应用的可行性[11]。这些研究均验证了知识图谱与文化遗产研究结合的可行性，为本研究提供方法上的指导。

1.2 基于知识图谱的智能问答研究

智能问答的目的是从自然语言描述中得到准确的答案，实现人机交互。知识图谱作为语义网络，自提出之日起就被称为搜索引擎的知识库[12]，知识图谱被广泛应用于个性化推荐、智慧搜索和智能问答等领域，尤其是在智能问答领域得到研究人员的广泛关注。目前，基于知识图谱的智能问答主要有以下 3 种实现方法：①基于模板的问答方法，通过预先定义的规则或模板，将问题转换成查询语句，返回正确答案。例如丁斌通过模板库构建汽车领域的智能问答系统。该方法的优点是简洁，准确度高，但需要花费大量的人力来构建模板[13]。②基于语义解析的问答方法，主要是解析自然语言问句成分，并用查询语言对其进行映射，再通过知识图谱查询相关答案。如

高劲松等在馆藏文物资源关联数据模型的基础上，构建馆藏文物资源智能问答系统[14]。该方法可解释性较强，在领域内的应用效果较好，但缺乏通用性。③基于深度学习的问答方法，是将自然语言问句表示为向量，利用深度学习模型计算向量间的相似度，对候选项排序，并给用户返回相似度最大的答案。如乔凯等引入关键词注意力机制构建中文医疗问答匹配系统[15]。该方法可应用于大规模数据集，但其可解释性较差，模型的训练成本高。这些方法为本研究构建智能问答系统提供了参考。

综上所述，在数字人文的背景下对黄河流域非物质文化遗产的研究还不多见，已有研究主要是对黄河流域历史文献的研究，较少考虑将语义网技术应用到黄河流域非物质文化遗产资源的知识组织与知识发现中。本研究通过图数据库 Neo4j 构建黄河流域非遗资源知识图谱，以图的形式存储黄河流域非遗资源，实现知识之间的互联，最后，设计了基于前后端交互的黄河流域非遗资源智能问答系统，以满足普通用户的查询与检索需求。

2　研究框架设计

本研究根据黄河流域非遗资源的特点，对其进行知识梳理，并参考领域知识图谱构建的流程，设计了黄河流域非遗资源智能问答研究框架，如图 1 所示。

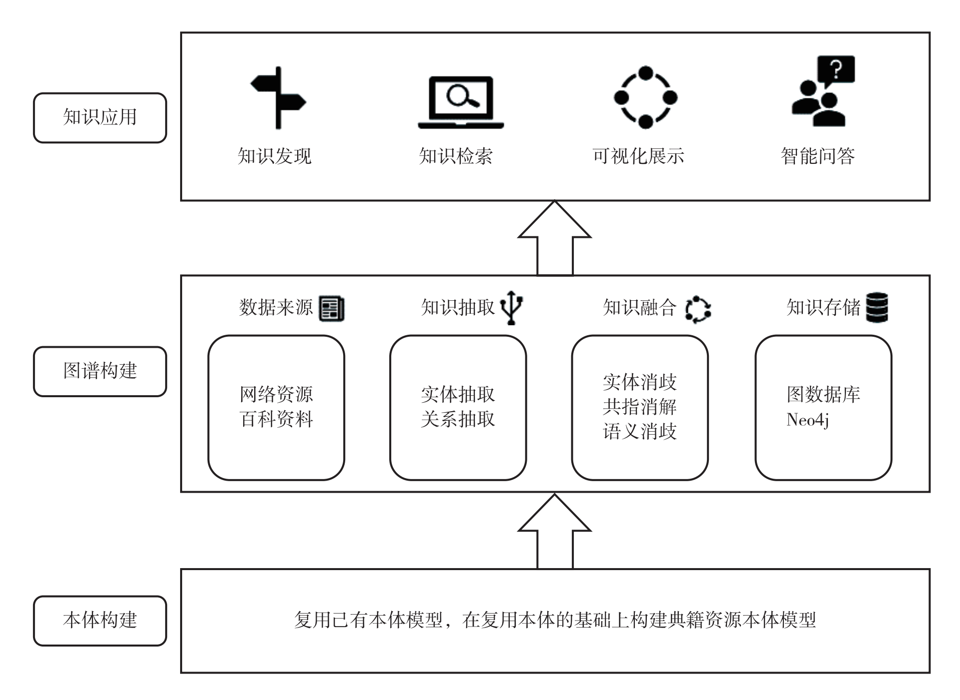


图 1　黄河流域非遗资源智能问答研究框架

2.1　本体构建与数据来源

黄河流域非遗资源作为一种领域性较强的研究对象，需要采用自顶向下的方式来构建知识图谱。本体作为构建知识图谱的核心，其设计的好坏直接关系到知识图谱的质量。因此，必须明确

黄河流域非遗本体的核心概念及所具有的属性，从而避免概念之间的歧义，并准确揭示出概念之间的语义关系。在构建本体前，需要参考已有的成熟本体，通过复用现有本体词表，并结合自建本体词表实现本研究的本体构建，以便满足黄河流域非遗资源知识组织与知识描述的需要。在确定黄河流域非遗资源的本体模型之后还需要根据数据来源，结合数据的特点改进已构建的本体模型。数据来源主要包括中国非物质文化遗产网、黄河非遗数据库、各省非物质文化遗产网及网络百科资源等。

2.2 知识抽取

知识抽取是指根据已构建的本体模型，从多源异构的数据中将所需要的知识抽取出来。为保证知识抽取的准确性，应根据数据的来源和结构的不同，利用网络爬虫、模式匹配、机器学习等信息抽取技术，完成实体识别、属性抽取和关系抽取。实体抽取是对具有特定意义的实体进行抽取，主要包括非遗项目、人物名称、出生日期、涉及事件和代表作品等信息；属性抽取是将非遗项目、人物、事件、时间、地点、资源等概念类的描述信息抽取出来；关系抽取是获取实体与实体及实体与属性的语义关系并以三元组的形式表示出来。本研究通过神经网络与规则相结合的方式实现知识抽取，以三元组的方式将数据存储起来，为知识图谱的构建奠定基础。

2.3 知识融合

知识抽取完成之后，抽取出来的知识可能存在重复、异名同义的问题，需要通过共指消解和语义消歧等知识融合方法对其进行整合处理。共指消解是指某些实体或关系可能具有多种表达方式，如某些非遗项目在不同来源数据中的名称存在差异，但所指的实体是一致的，也存在不同的非遗项目所用的名称是一样的。除此之外，时间实体和关系也存在共指问题，如公元 1662~1722 年和清代康熙年间为同一时间，事件的发生时间和举办时间的含义是一样的。本研究通过人工和文本相似度计算相结合的方法来消除歧义，使用人工的方式来处理复杂的、领域性强的实体歧义，使用文本相似度计算将关系中相似度高于一定阈值的关系进行合并。

2.4 知识存储

知识融合任务完成之后，需要将实体、属性及关系存储起来，并以 RDF（Resource Description Framework）框架的方式表示。RDF 的基本数据单元是一个三元组，可以表示为 SPO（Subject-Predicate-Object）。使用 RDF 三元组资源描述框架可以揭示知识与知识之间的关系，其序列化方式主要有 RDF/XML、N-Triples、Turtle、RDFa、JSON-LD 等。与其他序列化方式相比，JSON-LD 易于解析与理解，更适合 Web 应用程序的开发，考虑到本研究在知识应用层需构建智能问答的交互系统，故采用 JSON-LD 以键值对的方式来有效地存储 RDF 三元组数据，通过调用 Python 语言的 py2neo 包连接图数据库 Neo4j 将 JSON 数据导入到图数据库中。

2.5 智能问答系统

完成知识图谱构建之后，在此基础上构建黄河流域非遗资源的智能问答系统。本系统主要包括两大部分：一是关于黄河流域非遗资源的知识检索，通过名称匹配的方式返回与之相关的知

识。二是关于黄河流域非遗资源的智能问答，在本研究中通过自然语言处理技术完成问句的分词与句法分析工作，解析出问句中的实体与关系，然后通过 py2neo 库在 Neo4j 中查询结果，智能问答系统会返回目标实体的图谱知识及其相关的信息。

3　实证研究：黄河流域非遗资源的智能问答

3.1　对象选取

本研究选取黄河流域非物质文化遗产资源为对象进行实证研究。黄河流域非遗资源具有三大特征：①资源相对分散。黄河流域非遗资源主要分布在中国非物质文化遗产网上，但部分数据存在缺失的情况，比如：师承关系、图片视频资源等，还需通过各省份的非遗网站获取相关数据进行统一的整理与组织。②关联关系多样。黄河流域非遗资源之间蕴含着复杂多样的关联关系，如非遗之间的相关关系，人物之间的师承关系等，为黄河流域非遗资源间的知识关联奠定基础。③语义内容丰富，黄河流域非遗资源具有丰富的人物、时间、地点、事件和资源等要素，要素与要素之间、同一要素之间都蕴含着复杂丰富的语义内容，需要对黄河流域非遗资源进行语义解析与组织。此外，黄河作为水文化遗产，是中华文明最主要的发源地，以黄河流域非遗资源作为研究的对象，在理论与实践方面均有较高的研究价值。

3.2　本体构建

构建黄河流域非遗本体模型需要对非遗本体信息、人物、时间、地点、事件和资源等语义信息进行描述，现有可利用的本体模型具有一定的借鉴意义。因此，在构建本体时，可参考 CRM（Conceptual Reference Model）文化遗产领域通用模型[16]、DC（Dublin Core）[17]、社会网络本体 FOAF（Friend of a Friend）[18]、上海图书馆名人手稿档案库[19]等通用本体进行复用。基于此，本研究复用人物类（foaf: Person）、时间类（crm: TimeSpan）、地点类（crm: Place）与事件类（crm: Event），并以 yrich（Yellow River Intangible Cultural Heritage）作为自定义本体命名空间，命名了项目类（yrich: Project）和资源类（yrich: Resources），共六大类别来描述黄河流域非遗资源中涉及的知识类型。

3.2.1　核心类构建

基于上文分析，根据黄河流域非遗资源的语义关联需要，构建了黄河流域非遗本体的核心概念层级，共划分成以下六个核心类。

（1）项目类

项目类是本研究的研究主题，即黄河流域非物质文化遗产，非遗项目作为本体构建的核心类，与人物、时间、地点、事件和资源之间存在关联关系，项目之间也存在相关关系，如晋剧（内蒙古）和晋剧（太原）的申报地区不一样，但都是晋剧在黄河流域的发展分支，因此，它们之间存在相关关系。非遗项目的数据属性描述其核心信息，如：项目名称、项目编号、批次、公布时间、类别等。

（2）人物类

人物类涵盖了与非遗项目相关的传承人、研究者和相关者。传承者是传承和弘扬非物质文化

遗产的人，与非遗项目的发展和传承息息相关。研究者是指研究非遗发展现状的人，深入研究非遗的传播、利用与传承，为推动非遗的发展提出新方案。相关者是指非遗项目涉及的历史人物。人物类不仅与其他核心概念类存在关系，人物之间还存在师承关系。人物类的数据属性主要包含了姓名、人物类型、序号、性别和民族等。

（3）时间类

时间类是指非遗项目、人物、事件所记录的时间信息，根据时间构成方式的不同，分为两个子类，包括抽象时间类（crm: TimeAbstract）和具体时间类（crm: TimeSpecific）。抽象时间主要指无法具体到某一年的时间，如公元 8 世纪，具体时间是可以用年月日来表示的时间，如 1946 年 9 月 1 日。

（4）地点类

地点类包含了非遗项目、人物、事件所涉及的位置信息，如：非遗项目的申报地点与起源地、人物的居住地址及事件发生的地点。地点类与非遗项目、人物和事件之间存在关联，而其本身也具有数据属性，包括地点名称和经纬度信息。

（5）事件类

事件类是指黄河流域非遗资源中所记录的相关事件，是人物、场景、时间和地点等元素的综合体。因此，事件类与项目类、人物类、时间类和地点类之间存在对象属性关系，其数据属性有事件名称、事件类型和事件描述，事件的类型主要包括民俗活动、祭祀活动和节日等。

（6）资源类

资源类是指黄河流域非遗数据中所具有的资源，如图片资源、视频资源、作品、研究论文等。资源类的数据属性主要有资源名称、资源类型、资源描述和资源链接。

3.2.2 属性创建

明确黄河流域非遗本体核心概念层级之后，还需要定义和描述概念类的各种属性，属性包括对象属性和数据属性，对象属性用来描述类与类之间的关系，其定义域和值域都属于类；数据属性用来描述类目的信息，定义域为类，值域是数据类型。本研究中对象属性主要包括项目间关系、项目与人物、项目与时间、项目与地点、项目与事件、项目与资源、人物与人物、人物与时间、人物与地点、事件与人物等关系，其概念类的对象属性信息如表 1 所示。

表 1　黄河流域非遗本体对象属性表

关系	关系类型	对象属性	定义域	值域
项目间关系	相关项目	yrich: relatedProjects	yrich: Project	yrich: Project
项目与人物	传承者	yrich: inheritor	yrich: Project	foaf: Person
	研究者	yrich: researcher		
	相关者	yrich: relatedPeople		

续表

关系	关系类型	对象属性	定义域	值域
项目与时间	起源时间	yrich: originTime	yrich: Project	crm: TimeSpan
	形成时间	yrich: formationTime		
项目与地点	申报地点	yrich: hasDeclarationPlace	yrich: Project	crm: Place
	起源地	yrich: hasOriginPlace		
项目与事件	涉及事件	yrich: relatedEvents	yrich: Project	crm: Event
项目与资源	拥有资源	yrich: hasResources	yrich: Project	yrich: Resources
人物与人物	师承	yrich: inheritFrom	foaf: Person	foaf: Person
人物与时间	出生日期	foaf: birthDay	foaf: Person	crm: TimeSpan
	死亡日期	foaf: deathDay		
人物与地点	居住地	yrich: residence	foaf: Person	crm: Place
人物与资源	成果	yrich: hasWork	foaf: Person	yrich: Resources
事件与人物	涉及人物	yrich: involvedPeople	crm: Event	foaf: Person
事件与时间	发生时间	yrich: tookPlaceAt	crm: Event	crm: TimeSpan
事件与地点	发生地点	yrich: occurrenceTime	crm: Event	crm: Place

类与类之间除了存在对象关系以外，还需要数据属性来表达其自身特征，黄河流域非遗本体的数据属性含义和值域如表 2 所示。

表 2　黄河流域非遗本体数据属性表

类	含义	数据属性	定义域	值域
项目	项目名称	dc: title	yrich: Project	xsd: string
	项目编号	yrich: number		
	批次	yrich: batch		
	公布时间	yrich: date		
	类别	yrich: heritageCategory		
	类型	dc: type		
	保护单位	yrich: contributer		
	描述	dc: description		

续表

类	含义	数据属性	定义域	值域
人物	姓名	foaf: name	foaf: Person	xsd: string
	人物类型	yrich: personType		
	序号	yrich: serialNo		
	性别	foaf: gender		
	民族	yrich: nation		
	所属机构	foaf: organization		
	简介	shl: briefBiography		
时间	时间值	yrich: timeValue	crm: TimeSpan	xsd: dateTime
地点	地点名称	yrich: placeName	crm: Place	xsd: string
	经纬度	yrich: longitudeAndLatitude		
事件	事件名称	yrich: eventName	crm: Event	xsd: string
	事件类型	yrich: eventType		
	事件描述	yrich: eventInfo		
资源	资源名称	yrich: resourcesName	yrich: Resources	xsd: string
	资源类型	yrich: resourcesType		
	资源描述	yrich: resourcesInfo		
	资源链接	yrich: resourcesURL		xsd: anyURI

本研究在完成黄河流域非遗本体模型构建之后，明确定义了概念类及其相关属性，将黄河流域非遗资源中的项目、人物、时间、地点、事件和资源等类添加到本体模型，黄河流域非遗本体模型共包含 6 大类、原有的 2 个子类，本意是想表达时间类下还有抽象时间与具体时间 2 个子类，删除不影响表达。13 个对象属性和 25 个数据属性，最终设计的黄河流域非遗资源本体模型如图 2 所示。

3.3　知识抽取与融合

3.3.1　数据获取

本研究所需要的数据来自中国非物质文化遗产网、各省非物质文化遗产网及网络百科资源等，通过自编 Python 爬虫程序来爬取中国非物质文化遗产网关于黄河流域九大省份非遗的相关数据，为使内容更加的完整全面，还从百度百科及各省非物质文化遗产网上获取到相关数据，以完善非遗项目的相关信息。此外，爬取百度百科上的非遗、人物和资源的图像和视频作为黄河流域非遗资源的多模态数据，图像等资源则是以其描述的实体名称进行命名保存，以便后续网页系统的可视化呈现。

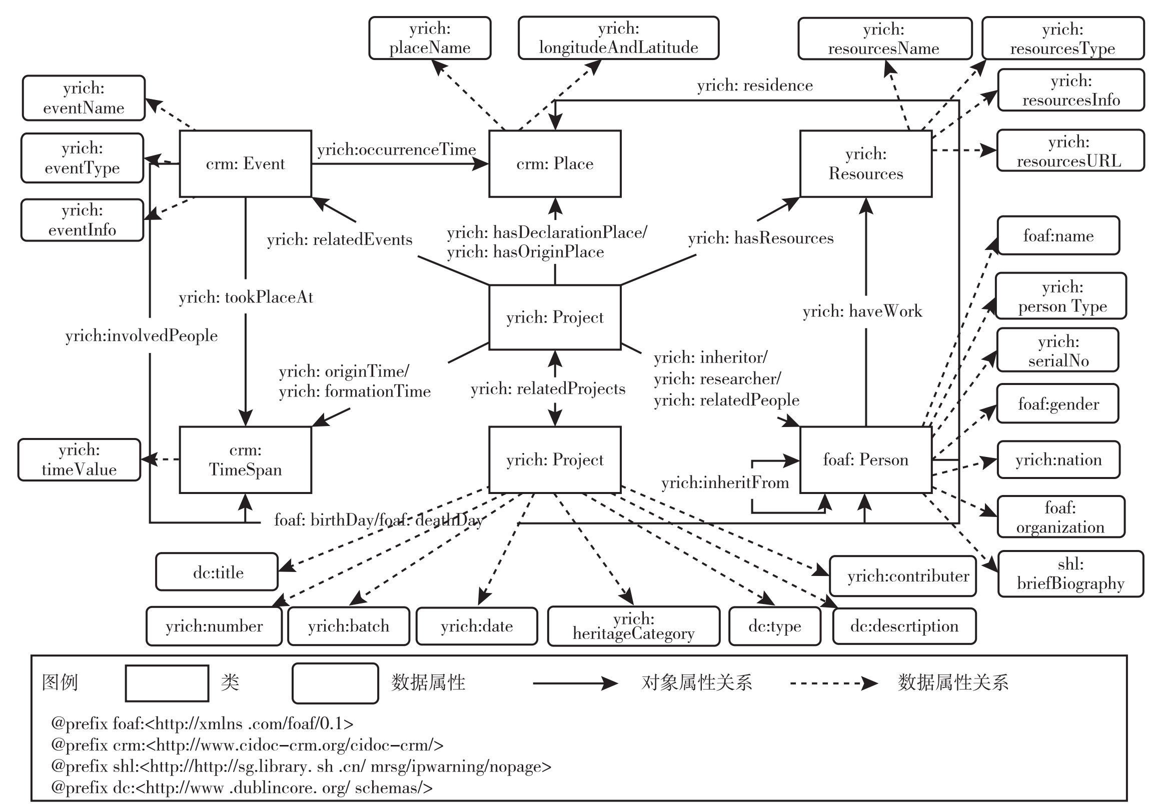


图 2 黄河流域非遗资源本体模型

3.3.2 知识抽取

知识抽取是知识图谱构建的关键一环，其过程是在已定义好的本体模型基础上，抽取所需的实体、属性和关系等知识。当前的知识抽取主要指的是关系抽取，原因在于关系抽取的结果会以三元组的形式表示出来，一并抽取出了实体及属性值。本文根据数据结构的不同分别采用了不同的方法进行知识抽取。针对非遗网站的文本数据，采用了基于句法规则的三元组关系抽取，通过调用 HanLP[20]，对爬取到的网页文本进行词法分析、句法分析与语义分析，抽取三元组关系。针对网络百科资源，采用了基于神经网络的三元组关系抽取，通过调用 Jiagu 深度学习自然语言处理工具来实现关系抽取，其以 BiLSTM 等模型为基础，使用大规模中文语料训练而成，所使用的语料大多数来源于百度百科。例如以“白秉权，1930 年生于陕西省西安市”为输入语句，两种模型分别抽取句子中的三元组关系，结果如图 3、图 4 所示。可见，利用 Jiagu 对百科资源进行抽取的效果较好。

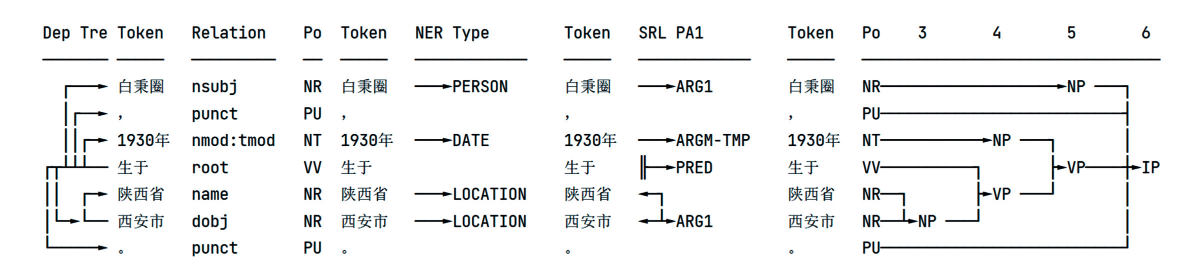


图 3 HanLP 三元组抽取

```
import jiagu
text = '白秉权，1930年生于陕西省西安市'
words = jiagu.cut(text)
# print(words)
knowledge = jiagu.knowledge(text)
print(knowledge)
```

```
三元组抽取 ×
D:\anaconda3\envs\py3.8\python.exe D:/pythonProject/jiagu/三元组抽取.py
[['白秉权', '出生日期', '1930'], ['白秉权', '出生地', '陕西省西安市']]

进程已结束,退出代码0
```

图 4　Jiagu 神经网络模型三元组抽取

3.3.3　知识融合

多源异构的黄河流域非遗数据经过知识抽取后，形成包含实体和关系的三元组数据集，但部分实体和关系仍然存在表达冗余和语义歧义的问题。由于本研究的实体消歧问题复杂且领域性强，无法通过常规的算法对其融合，为解决这一问题，本文采用人工构建自定义词典完成实体融合。针对关系间存在的歧义问题，选择皮尔逊相似系数结合哈工大同义词词林（扩展版）来计算关系名的相似度，首先将关系名转换成词向量，通过皮尔逊相似系数公式计算得出向量之间的相似度值 $P(x, y)$，其计算公式如公式（1）所示：

$$P_{(x,\ y)} = \frac{\sum(X-\bar{X})\ (Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \tag{1}$$

P 的取值范围在［-1，1］之间。参考已有研究，一般认为值超过 0.8 就代表 x、y 存在极强相关，可以认为两个关系名称属于同一关系，即可进行关系的合并，反之，x、y 不属于同一关系。

3.4　知识存储

完成知识融合任务之后，本文选择 Neo4j 图数据库来实现知识存储。Neo4j 作为一种开源、可伸缩的高性能图形数据库，与传统的 SQL 数据库相比，可以更好地表示和处理关系数据。Neo4j 是由标签、节点、属性和关系组成的，用来存储具有关联关系的数据，实现对多源异构数据的知识描述与语义组织，为后续的知识关联、知识查询和知识问答提供支持。

本文使用 Neo4j 桌面版将汇总整理的三元组数据存储起来，首先通过调用 Python 的第三方库 py2neo，利用 Cypher 语句将三元组数据导入到 Neo4j 中。本文共构建了 4 442 个实体节点和 7 487 条三元组关系，其部分的可视化界面如图 5 所示。

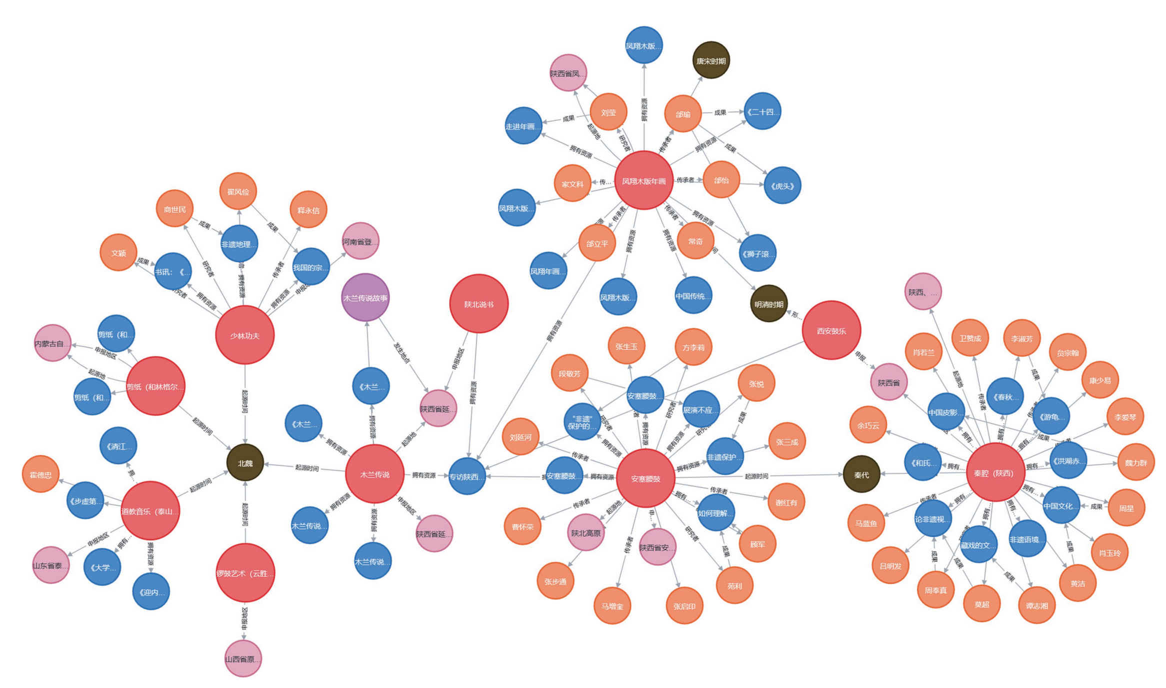


图5 黄河流域非遗资源知识图谱（部分）

3.5 智能问答

本研究搭建了黄河流域非遗智能问答原型系统，以实现对黄河流域非遗资源的知识交互。智能问答原型系统采用前后端分离技术，前端通过 HTML、CSS、JavaScript 和 Bootstrap 来完成对网页端的页面布局和用户交互，后端采用 Python 语言通过统一的接口去调用 Neo4j 图数据库，实现智能问答系统的知识生成、检索和问答等应用。在前后端交互上采用了 Flask Web 框架，Flask 是基于 Python 语言的轻量级 Web 应用框架，与其它的框架相比，其灵活、易上手、可扩展性强，符合本研究的系统构建需求。

智能问答系统主要提供知识图谱展示、知识检索、知识问答三大功能。知识图谱展示模块展示了黄河流域 619 个非遗项目的关系全貌图，其实现的过程是通过 Echarts 的力导向图将 JSON 格式的数据在前端呈现出来，便于用户查看黄河非遗的关系全貌。其与后端的 Neo4j 数据库呈现的知识全貌仅是外观上有所区别，其本质无异。本节主要介绍知识检索和知识问答两大功能。

3.5.1 知识检索

黄河流域非遗资源知识分散，关系检索可以关联检索到与非遗项目相关的知识，如传承者、起源时间、申报地区、相关事件与资源等，实现黄河流域非遗资源的知识重组与语义关联。关系检索的原理是后端接收到检索框中输入内容，通过调用 py2neo 包将检索结果从 Neo4j 图数据库中返回，然后将查询结果即节点与关系数据存储为 JSON 格式文件，利用 Echarts 中的力导向图在前端展示出来。以陕西省非遗项目安塞腰鼓为例，展示实体之间的关联关系，检索结果如图 6 所示。

图 6　黄河流域非遗项目关系检索

实体在前端展示时主要是通过地区的不同加以区分，非遗项目根据其所属省份呈现相应的颜色，人物根据其居住地址确定其所属地区，事件以发生地点为依据，资源与其所相关联的非遗项目为标准，时间实体没有地点属性，其颜色划分为其他。

3.5.2　知识问答

智能问答系统的目的是为了实现人机交互，通过解析用户输入的自然语言，将答案返回给用户。本研究设计了基于语义解析的智能问答系统，主要包括自然语言处理、知识图谱查询和图谱生成三个部分。自然语言处理是将文本数据转换成可被机器理解的语言，问答系统接收到用户在前端输入的问题之后，调用 HanLP 分词工具对问句进行自动分词、词性标注和去除停用词等处理，识别出问句中包含的实体和关系，如：非遗项目、人名等实体名及起源时间、起源地等关系名。为保证自动分词的准确度，通过自定义词典和哈工大的同义词词典来处理实体和关系歧义问题。知识图谱查询是根据问句处理之后生成的查询语句到 Neo4j 图数据库匹配目标实体，问答系统将获取到的数据以 JSON 格式的数据返回给前端生成知识图谱，同时返回实体对应的图片资源和基本信息，其中图片资源依据实体名称进行匹配。

以“川剧的传承者？”这一问题为例，当问句返回到后端之后，首先，利用 HanLP 对

其进行解析，同时利用自定义词典和同义词词典对其进行约束，解析出实体“川剧”、关系“传承者”，之后到后端查询结果，其查询语句为“MATCH (n: 非遗项目)-[r: ‘传承者’]->(p) WHERE n.name= ‘川剧’ return n, r, p”。最后，在前端展示出川剧所有的传承者及关系，针对有图片资源的节点还会返回对应的图片信息，实现了多模态的信息检索，结果如图 7 所示。

图 7　黄河流域非遗资源智能问答系统示例

3.5.3　系统测评

问答系统的性能评价是了解一个问答系统的可行性和实用性的关键所在。本研究的测试环境和软件版本主要为：Ubuntu20.04 操作系统、Python3.6、Neo4j_Desktop1.5.8、Flask2.0.3 等。为验证本研究构建的基于知识图谱的黄河流域非遗资源智能问答系统的性能，选取准确率（Precision）作为评价指标，如公式（2）所示。

$$Precision = \frac{x}{y} \times 100\% \tag{2}$$

其中，x 代表正确回答问题的个数，y 代表输入的问题总数。本研究通过问卷采集了黄河流域非遗资源的热门问题共 500 余条，经过人工校对、分类标注后制作为包括非遗基本信息、人物社会关系、非遗相关事件、非遗相关资源四类问题，共 400 个问题测试对，每类问题 100 条。经过实验测试，回答准确率达到了 91.5%，验证了本系统的有效性和准确性，基本可以满足普通用户的问答需求，具体的实验结果如表 3 所示。

表 3　智能问答系统性能测试表

序号	问题类型	测评问题数（个）	准确回答数（个）	回答准确率（%）
1	非遗基本信息	100	94	94
2	人物社会关系	100	95	95
3	非遗相关事件	100	89	89
4	非遗相关资源	100	88	88
	总计	400	366	91.5

由表 3 可以看出，本研究的问答系统对非遗基本信息、人物社会关系类问题的回答准确率较高，对非遗相关事件及非遗相关资源类问题的回答准确率还需加强，主要原因在于非遗资源的基本信息及人物社会关系的解析较为明晰，相关节点与关系名称匹配度较高，不易发生歧义。而在非遗相关事件及非遗相关资源类问题上，问句的方式往往差异化较大，节点名词与用户的提问之间未能很好的转换识别，造成在解析节点名称上难以与自定义的词典进行匹配，进而影响回答的效果。

4　结　语

本文立足于数字人文视角，以黄河流域九大省份的非物质文化遗产资源作为数据基础，通过自顶向下的方式构建了黄河流域非遗资源知识图谱，并设计了基于知识图谱的黄河流域非遗资源智能问答原型系统，为黄河流域非遗资源的数字化建设提供了新的思路与方法。本系统基本实现了对黄河流域九大省份非遗资源的深度挖掘和揭示，同时也可为其他非遗资源保护和传承提供数字化的借鉴思路与支持。

本研究的理论与实际价值在于：（1）以知识图谱的形式重新将黄河流域非遗资源的相关知识进行关联，使离散的非遗资源以图的方式串联起来，有助于黄河流域非遗资源的可视化呈现。（2）在智能问答构建上，采用了基于语义解析的问答方式，既保证了解析的准确率，又避免了深度学习需要大规模数据的弊端。（3）智能问答系统采用了前后端交互的方式，使用户仅需采用自然语言即可进行检索查询。在未来的工作中，本研究将完善前后端的交互效率，使之可以运行在高负荷高并发的互联网环境中。

【参考文献】

［1］习近平. 在黄河流域生态保护和高质量发展座谈会上的讲话［J］. 中国水利，2019（20）：1-3.

［2］中国政府网. 中共中央办公厅 国务院办公厅印发《关于进一步加强非物质文化遗产保护工作的意见》［EB/OL］.［2022-10-31］. http://www.gov.cn/zhengce/2021-08/12/content_5630974.htm.

［3］Luhmann J, Burghardt M. Digital humanities—A discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape［J］. Journal of the Association for Information Science and Technology,

2022, 73(2): 148–171.

［4］肖鹏，姚楚晖. 走向数字时代的人文学者：泛LIS视域下的进展与反思［J］. 文献与数据学报，2019，1（3）：18–25.

［5］张强，高劲松，龙家庆，等. 基于知识重构的词人时空情感轨迹可视化研究——以辛弃疾为例［J］. 情报学报，2023，42（6）：729–739.

［6］刘峤，李杨，段宏，等. 知识图谱构建技术综述［J］. 计算机研究与发展，2016，53（3）：582–600.

［7］高劲松，张强，李帅珂，等. 数字人文视域下诗人的时空情感轨迹研究——以李白为例［J］. 数据分析与知识发现，2022（9）：1–19.

［8］赵雪芹，李天娥，曾刚. 基于Neo4j的万里茶道数字资源知识图谱构建研究［J］. 情报资料工作 2022，43（5）：89–97.

［9］高劲松，张强，李帅珂. 可移动文物的知识图谱构建及关联数据存储——以湖北省博物馆为例［J］. 现代情报，2022，42（4）：88–98.

［10］范青，史中超，谈国新. 非物质文化遗产的知识图谱构建［J］. 图书馆论坛，2021，41（10）：100–109.

［11］赵雪芹，路鑫雯，李天娥，等. 领域知识图谱在非遗档案资源知识组织中的应用探索［J］. 档案学通讯，2021（3）：55–62.

［12］Singhal A. Introducing the knowledge graph: things, not strings［J］. Official google blog, 2012, 5(16): 3.

［13］丁斌. 汽车领域智能问答系统中模板库自动生成方法的研究［D］. 上海：上海交通大学，2017.

［14］高劲松，方晓印，刘思洋，等. 基于关联数据的馆藏文物资源知识关联与智能问答研究［J］. 情报科学，2021，39（5）：12–20.

［15］乔凯，陈可佳，陈景强. 基于知识图谱与关键词注意机制的中文医疗问答匹配方法［J］. 模式识别与人工智能，2021，34（8）：733–741.

［16］CIDOC CRM［EB/OL］.［2022–10–30］. https://cidoc–crm.org/.

［17］DCMI Schemas［EB/OL］.［2022–10–30］. https://www.dublincore.org/schemas/.

［18］FOAF［EB/OL］.［2022–10–30］. http://xmlns.com/foaf/.

［19］上海图书馆名人手稿［EB/OL］.［2022–10–30］. http://sg.library.sh.cn/mrsg/ipwarning/nopage.

［20］HanLP［EB/OL］.［2022–12–17］. https://www.hanlp.com/.

张强，吴艳飞，高颖，等 . 基于知识图谱的黄河流域非遗资源智能问答研究［J］. 文献与数据学报，2023，5（3）：100-115.

Research on Intelligent Question and Answer of Intangible Cultural Heritage Resources in the Yellow River Basin Based on Knowledge Graph

Zhang Qiang[1,2,3]　Wu Yanfei[1]　Gao Ying[1,2]　Zhou Shubin[1]

(1.School of Information Management, Central China Normal University, Wuhan 430079, China;
2.Institute of Digital Humanities, Renmin University of China, Beijing 100872, China;
3.School of Journalism and Communication, Anhui Normal University, Wuhu 241002, China)

Abstract: ［**Purpose/significance**］The use of digital technology to empower the in-depth excavation of intangible cultural heritage resources and reveal the correlation between intangible cultural heritage resources in the Yellow River Basin is of great significance to the protection and inheritance of intangible cultural heritage in the Yellow River Basin. ［**Method/process**］Taking the intangible cultural heritage resources of the Yellow River Basin as the research object, the knowledge graph of the intangible cultural heritage resources in the Yellow River Basin was constructed in a top-down manner, and an intelligent question and answer system with user interaction as the core was constructed. ［**Result/conclusion**］The intelligent question-and-answer system of intangible cultural heritage resources in the Yellow River Basin based on knowledge graph constructed in this study realizes the multi-dimensional knowledge discovery of intangible cultural heritage resources in the Yellow River Basin, and provides new ideas for the related research of intangible cultural heritage resources.

Keywords: Knowledge graph; Intangible cultural heritage; Yellow River basin; Intelligent question and answer

（本文责编：王秀玲）